EuPA OPEN PROTEOMICS 7 (2015) 11–19

ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

Detecting significant changes in protein abundance

Kai Kammers[a], *Robert N. Cole*[b], *Calvin Tiengwe*[c,d], *Ingo Ruczinski*[a,*]

[a] *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA*
[b] *Mass Spectrometry and Proteomics Core Facility, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[c] *Department of Cell Biology, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[d] *Department of Microbiology and Immunology, School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA*

ARTICLE INFO

Article history:
Available online 25 February 2015

Keywords:
Empirical Bayes
Inference
Protein abundance

ABSTRACT

We review and demonstrate how an empirical Bayes method, shrinking a protein's sample variance towards a pooled estimate, leads to far more powerful and stable inference to detect significant changes in protein abundance compared to ordinary t-tests. Using examples from isobaric mass labelled proteomic experiments we show how to analyze data from multiple experiments simultaneously, and discuss the effects of missing data on the inference. We also present easy to use open source software for normalization of mass spectrometry data and inference based on moderated test statistics.

1. Introduction

Detecting significant changes in protein abundance is a fundamental task in mass-spectrometry based experiments when trying to compare treated to untreated cells, wildtypes to mutants, or samples from diseased to non-diseased subjects. The statistical inference for proteomic data in these settings is usually based on standard 2-sample t-tests, comparing the measured relative or absolute abundances for each peptide or protein across the conditions of interest. However, sample sizes are often small, sometimes as small as 4 or 8 samples total, which result in great uncertainty in the sample variability estimates. Since these estimates are used in the test statistics to assess the statistical significance of the observed fold change, proteins exhibiting a large fold change are often declared non-significant because of a large sample variance, while at the same time small observed fold changes might be declared statistically significant, because of a small sample variance.

Additional methods to assess biological and technical sources of variability have been proposed [1–6], including methods to analyze data from multiple experiments simultaneously. For case–control iTRAQ experiments, Oberg et al. [7] and Hill et al. [8] extended a linear mixed effects approach originally proposed by Kerr and Churchill [9,10] as analysis of variance for gene expression studies. This mixed model adjusts for potential differences due to channel effects, loading, mixing, and sample handling. The parameter of interest in the model is the interaction between protein and group status, with a statistically significant result indicating differential expression (abundances) between cases and controls. One of the noteworthy features of this approach is that it simultaneously estimates protein relative abundance and assesses differential expression, albeit with substantial computational cost due to the numerical complexity of optimizing the likelihood and estimating a rather large number of parameters. Herbrich et al. [11] demonstrated that estimating protein abundances using median sweeps reduces computational cost substantially, and is as efficient yet more robust than

* *Corresponding author*. Tel.: +1 4106147840.
E-mail address: ingo@jhu.edu (I. Ruczinski).
http://dx.doi.org/10.1016/j.euprot.2015.02.002

protein abundance estimation procedures based on linear mixed effects models.

An implicit assumption in the approach of Oberg et al. [7] and Hill et al. [8] is that the biological variability is the same for all proteins identified and quantified. Though *"all models are wrong, but some are useful"* [12], incorrect model assumptions can lead to a loss in power even if no bias is incurred. This was for example observed in gene expression studies when LIMMA ("Linear Models for Microarray Data") [13] was introduced as an empirical Bayes approach that specifically allowed for a realistic distribution of biological variances, compared to the models of Kerr and Churchill [9,10], which assumed constant variability. The statistical trick in LIMMA is to use the full data to shrink the observed sample variances towards a pooled estimate. This results in far more stable and powerful inference compared to ordinary *t*-tests particularly when the number of samples is small [13], yet still allows for a distribution of variances. LIMMA arguably is the contemporary analytical standard for gene expression experiments, as evidenced by over 6000 citations in the last ten years (http://scholar.google.com). LIMMA has also been sporadically used in the context of proteomic experiments [14–19], but is far from being regarded as the analytical standard. This is surprising since proteomic experiments often have somewhat small sample sizes, and for those the potential gains of an empirical Bayes procedure are highest. One possible explanation for this phenomenon (besides being originally developed for a different genomic application) might be that LIMMA has been implemented as a Bioconductor package in the language R, a statistical environment the proteomics community only recently started to embrace [20–26].

In this manuscript we use examples from quantitative proteomic experiments using isobaric mass tags to demonstrate how better results in case–control studies can be achieved by using the LIMMA moderated test statistics. We show how to analyze data from multiple experiments simultaneously, and discuss the effects of missing data on the inference. We give sufficient detail for the statistically inclined reader to understand what happens "under the hood" of this empirical Bayes approach, and also present easy to use open source software for the practitioner to carry out the normalization of these mass spectrometry data, and to readily obtain the inference from moderated test statistics.

2. Materials and methods

2.1. Sample description

The data stem from two *Trypanosoma brucei* transgenic cell lines overexpressing either *TbHslV*-wild type or *TbHslV*-mutant protease. The *T. brucei* mitochondrion contains a proteasome-like ATP-dependent protease named *TbHslVU* that plays a critical role in regulating the timing of mitochondrial DNA replication [27]. Previous experiments suggested that *TbHslVU* controls the timing of kDNA synthesis by degrading "positive regulator of replication" [27,28]. To search for *TbHslVU* substrates its catalytically active subunit (denoted as *TbHslV-wt*) and its catalytically dead mutant (denoted as *TbHslV-mt*) were fused to the tandem affinity purification (TAP) tag. TAP-tagged *TbHslV-wt* or *TbHslV-mt* overexpressing cell lines were generated and the overexpressed proteins were purified using a TAP protocol adapted from Ringpis [29]. *TbHslV-wt* and *TbHslV-mt* were performed in four independent biological replicates.

Quantitative mass spectrometry was used to identify proteins that are associated with overexpressed and purified *TbHslV-mt* but not with *TbHslV-wt* treated similarly; since the latter binds and degrades its substrates. Proteins were digested with trypsin, labelled using the eight-plex iTRAQ isobaric mass tags (ABSciex) and analyzed using tandem mass spectrometry on an LTQ Velos Orbitrap interfaced with an Eksigent 2D NanoLC as previously described [11,30,31], except mass tagged peptides were fractionated by basic reverse phase chromatography [32]. Peptides were identified using Proteome Discoverer v1.4 (Thermo Scientific, San Jose, CA) and Mascot v2.2 (Matrix Sciences). Software defaults were used to control the false discovery rate (FDR) and only peptides spectra with less than 1% FDR and less than 30% isolation interference were included in analysis.

Protein $\log_2$ relative abundances were estimated using the method of Herbrich et al. [11]. In this procedure, a logarithmic transformation of the reporter ion intensities is employed since systematic effects and variance components are usually assumed to be additive on this scale [7,8]. The $\log_2$ reporter ion intensities for each spectrum are "median-polished" by subtracting the spectrum median $\log_2$ intensity from the observed $\log_2$ intensities. The relative abundance estimate for a particular protein is calculated as the median of these residuals, from all reporter ion intensity spectra belonging to this protein. Corrections for differences in amounts of material loaded in the channels and sample processing are carried out by subtracting the channel median from the relative abundance estimate, normalizing all channels to have median zero.

2.2. Statistical inference

2.2.1. Two group comparisons

To detect differentially expressed proteins in a balanced proteomic experiment with n cases ($\log_2$ relative abundances X_{1p}, …, X_{np} for protein p) and n controls ($\log_2$ relative abundances Y_{1p}, …, Y_{np}), inference is typically based on a 2-sample *t*-test for each protein p, with test statistic

$$t_p = \frac{\text{estimated log fold change}}{\text{estimated standard error}} = \frac{\overline{X}_p - \overline{Y}_p}{s_p\sqrt{2/n}}, \quad (1)$$

where $\overline{X}_p$ and $\overline{Y}_p$ are the group mean $\log_2$ relative abundances, and

$$s_p = \sqrt{\frac{\sum_i (X_{ip} - \overline{X}_p)^2 + \sum_i (Y_{ip} - \overline{Y}_p)^2}{2n-2}} \quad (2)$$

is the within-group sample standard deviation. For each protein, a p-value is then calculated referring the test statistic t_p to a *t*-distribution with $d_p = 2 \times n - 2$ degrees of freedom as null distribution. For the above the $\log_2$ relative abundances are assumed to be normally distributed with equal variance in each group, although *t*-tests are robust to departures from the

normality assumption unless outliers are present and sample sizes are small [33]. Similar test statistics can be calculated for non-equal group variances and unbalanced experiments.

2.2.2. *Moderated statistics*

The above approach estimates the variance s_p^2 and the standard error $s_p\sqrt{2/n}$ for each protein separately (Eqs. (1) and (2)), and does not use information (such as experimental precision) shared across all proteins. An alternative approach "Linear Models for Microarray Data" (LIMMA) [13], also applicable for mass-spectrometry based high throughput experiments, uses the fact that under a normality assumption for the $\log_2$ relative abundances the sample variance follows a scaled χ^2 distribution

$$s_p^2|\sigma_p^2 \sim \frac{\sigma_p^2}{d_p} \times \chi^2_{d_p}, \tag{3}$$

where σ_p^2 is the true (unknown) variance, and d_p are the degrees of freedom associated with the experiment. In contrast to the ordinary 2-sample *t*-test where σ_p^2 is regarded as a fixed (but unknown) parameter, LIMMA is an Empirical Bayes procedure where the protein variances are assumed to follow a scaled inverse χ^2 distribution

$$\frac{1}{\sigma_p^2} \sim \frac{1}{d_0 \times s_0^2} \times \chi^2_{d_0}. \tag{4}$$

The parameters d_0 and s_0^2 are estimated from the observed data via maximum likelihood. Using such as a scaled inverse χ^2 distribution for the protein variances implies that the set of protein sample variances s^2 follows a scaled *F* distribution [13], e.g.

$$s^2 \sim s_0^2 \times F_{d_p, d_0}. \tag{5}$$

Under the above hierarchical model, the posterior for a protein's sample variance is moderated: the observed protein sample variance s_p^2 is shrunk towards the common prior value s_0^2, with the magnitude of shrinkage depending on the relative sizes of the observed and prior degrees of freedom d_p and d_0:

$$s^2_{p\,[\text{moderated}]} = \frac{d_p \times s_p^2 + d_0 \times s_0^2}{d_p + d_0} = \lambda \times s_p^2 + (1-\lambda) \times s_0^2$$
$$\text{with} \quad \lambda = \frac{d_p}{d_p + d_0} \in (0, 1). \tag{6}$$

Thus, the shrinkage of the sample variance s_p^2 towards a common mean s_0^2 will be most pronounced when few data are available, as d_p and therefore λ will be small. The *p*-values are then derived referring the moderated *t*-statistic

$$t_{p\,[\text{moderated}]} = \frac{\text{estimated log foldchange}}{\text{moderated standard error}} = \frac{\overline{X}_p - \overline{Y}_p}{s_{p\,[\text{moderated}]}\sqrt{2/n}} \tag{7}$$

to a t distribution with $d_p + d_0$ degrees of freedom. Note that only the estimates of the standard errors change, and the estimated log fold changes in the numerator remain the same.

2.2.3. *Multiple experiments*

To achieve a desired sample size it is often necessary to carry out multiple experiments. These data can be analyzed by expanding the two-sample *t*-test into a more general linear model framework. For example, the regression for two 8-plex iTRAQ experiments can be written as

$$\mathrm{E}\left[Z_{pijk}\right] = \alpha_p + \tau_p \times \mathbf{1}_{\{i=1\}} + \gamma_p \times \mathbf{1}_{\{k=2\}} \tag{8}$$

where Z_{pijk} denotes the measured $\log_2$ relative abundances of protein p in sample $j \in \{1, \ldots, 4\}$ under condition $i \in \{0, 1\}$ in experiment $k \in \{1, 2\}$, i.e. 4 mutants and 4 wildtypes in each of two experiments (the $\mathbf{1}$ in Eq. (8) denotes the standard indicator function). For more than two experiments, the above equation can simply be expanded by allowing more parameters to indicate the extra experiments, or a mixed effects model can be used with a random effect for the experiment [7,8,11]. The inference to assess differential expression is on the parameter τ_p, the expected difference in $\log_2$ relative abundances of protein p when comparing a mutant and a wildtype from the same experiment. Statistical significance is based on the observed fold change, and an estimate of its standard error. The latter can be derived by estimating the variability separately for each protein. Alternatively, LIMMA also allows for pooling information across all proteins within the above linear model framework, generating moderated *t*-statistics and *p*-values [13].

2.2.4. *Multiple comparisons*

Since hundreds or even thousands of proteins can be identified and quantified in a typical mass spectrometry experiment, multiple comparisons corrections are imperative. The most popular procedure is the Bonferroni correction which controls the family wise error rate (FWER), that is, the probability of at least one type I error (i.e. false positive). Only proteins with nominal *p*-value less than α/N are declared differentially expressed, where α is the desired FWER (typically, 5%), and N is the number of proteins assessed. The consequence of such a strong protection against *any* false positives is usually a large number of false negatives, that is truly differentially expressed proteins not declared significant. Thus, in high-throughput experiments with potentially many differentially expressed proteins, a more desirable parameter arguably is the false discovery rate (FDR), designed to control the proportion of false positives among a set of proteins declared differentially expressed. The original FDR approach by Benjamini and Hochberg [34] was extended by Storey [35–37] to so-called "*q*-values", which have a similar interpretation for the FDR as *p*-values have for type-I error control: the *q*-value for a protein is defined as the minimum FDR that can be attained (i.e., the expected proportion of false positives incurred) when calling that protein differentially expressed. In other words: when testing for differential expression, if a protein has a *q*-value of 0.10, we expect 10% among the proteins that show smaller *p*-values to be false positives. The *q*-values are calculated from the observed *p*-values after estimating the proportion of differentially expressed proteins in the experiment (see Storey and Tibshirani [37] for details).

2.3. Simulations

To compare the performances of the ordinary 2-sample *t*-tests and the empirical Bayes moderated *t*-tests with regards to power and type I error (true and false positives), we simulated data mimicking the above described *T. brucei* 8-plex iTRAQ experiment with 4 mutants and 4 wildtypes. The $\log_2$ relative abundances of mutants and wildtypes for 1394 proteins were generated from a normal distribution with variances according to the scaled inverse χ^2 distribution (Eq. (4)) with estimated parameters $d_0 = 4.43$ prior degrees of freedom and a scaling factor of $s_0^2 = 0.032$ (Supplementary Fig. 1). The means of the normal distributions were chosen to reflect the desired fold changes in the respective simulations. A 50% fold change was used in all simulations when only a single experiment was considered, and both 25% and 50% fold changes were used in separate simulations when multiple experiments were analyzed.

2.4. Software

A software vignette to carry out the analyses described in this manuscript and to reproduce the simulations is available at www.biostat.jhsph.edu/~kkammers/software/eupa/. The software is freely available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.

3. Results

A total of 2047 proteins were identified and quantified in the *T. brucei* 8-plex iTRAQ experiment. To reduce the impact of possible false positive identifications, we only retained the 1394 proteins with 2 or more peptides quantified. With an ordinary 2-sample *t*-test to assess differential expression we find that most of the statistically significant differences were from proteins with little sample variability (Fig. 1A). Among the 1394 proteins, 258 (18.5%) achieved nominal significance at a 5% level, but only one protein was declared differentially expressed after multiple comparisons correction using the Bonferroni method (Fig. 1B). Notably, this protein had very small sample variability (8th smallest among the 1394 proteins). While only few more proteins were nominally significant at a 5% level when using moderated *t*-statistics (277, 19.9%), seven proteins achieved statistical significance after Bonferroni correction ($p < 3.6 \times 10^{-5}$, Fig. 1C). Moreover, at a FDR of 1% (i.e. $q < 0.01$) only 1 protein is declared differentially expressed when using ordinary test statistics, compared to 23 proteins declared differentially expressed when moderated test statistics are employed (FDR of 5%: 30 and 98 proteins, respectively; FDR of 10%: 120 and 184 proteins). Thus, at the same error level many more proteins can be declared differentially expressed when using the moderated compared to the ordinary *t*-statistics. This improvement is achieved by shrinking particularly the extreme variances towards a common mean (Supplementary Fig. 2). As a consequence, compared to their ordinary test statistics, proteins with low sample variability have less significant moderated test statistics, proteins with large sample variability have more significant moderated test statistics, and test statistics from proteins with moderate sample variability remain largely unchanged (Supplementary Fig. 3).

In a simulation study with ten differentially expressed proteins (50% fold change) we visualize how the differentially expressed proteins become more significant when moderated test statistics are used (Fig. 2A and B, green dots), while spurious associations with small observed fold changes tend to become less significant (Fig. 2A and B, yellow dots). This improvement in detecting differentially expressed proteins through moderated test statistics is not owed to an overall increase in false positives, as both the ordinary 2-sample *t*-test and the moderated *t*-test properly control the type I error (Fig. 2C and D). To quantify the potential gains in true discovery while maintaining error control via the FDR, we increased the number of differentially expressed proteins to 100 and averaged the results over 1000 iterations. For all levels of FDR control between 1% and 10%, the empirical Bayes approach using moderated test statistic produces substantially larger lists of proteins declared differentially expressed, detecting more truly differentially expressed proteins while maintaining proper error control (Fig. 3, left). Displaying the results on a receiver operating characteristic (ROC) curve, we find that at equal false positive rates the true positive rate is substantially higher when using moderated test statistics (Fig. 3, right).

Equivalent results were obtained when multiple experiments were analyzed simultaneously. We simulated three replicate 8-plex iTRAQ experiments with 100 differentially expressed proteins using the same parameters for the variability as above, and analyzed the data using the linear model displayed in Eq. (8). We used both ordinary test statistics (analyzing each protein separately) and the moderated test statistics using LIMMA. We recorded false positives and false negatives for a variety of significance thresholds, and averaged the results over 1000 iterations. Since the power to detect differentially expressed proteins increases with the number of experiments, we also show the findings for lower fold changes. For all false positive rates between 0.5% and 10% considered and fold changes of 25% and 50%, the true positive rate (sensitivity) was again higher when using the empirical Bayes approach with moderated test statistics (Fig. 4). In addition, we simulated experiments with three replicates and randomly selected 10% of proteins as missing in each experiment. While the presence of missing data generally results in a loss of power to detect differentially expressed proteins, these data can still be analyzed using linear models with moderated test statistics and proper type I error control. The findings for this setting were the same, supporting the variance shrinking approach as more powerful (Fig. 4). In particular, in the presence of missing data, the amount of shrinkage of the variance depends on the actual observed experimental degrees of freedom (Supplementary Fig. 4).

4. Discussion

Identifying differentially expressed proteins is a task in proteomic studies commonly carried out simply using *t*-tests. In this manuscript we demonstrated how better results can be achieved by using moderated *t*-statistics from the empirical

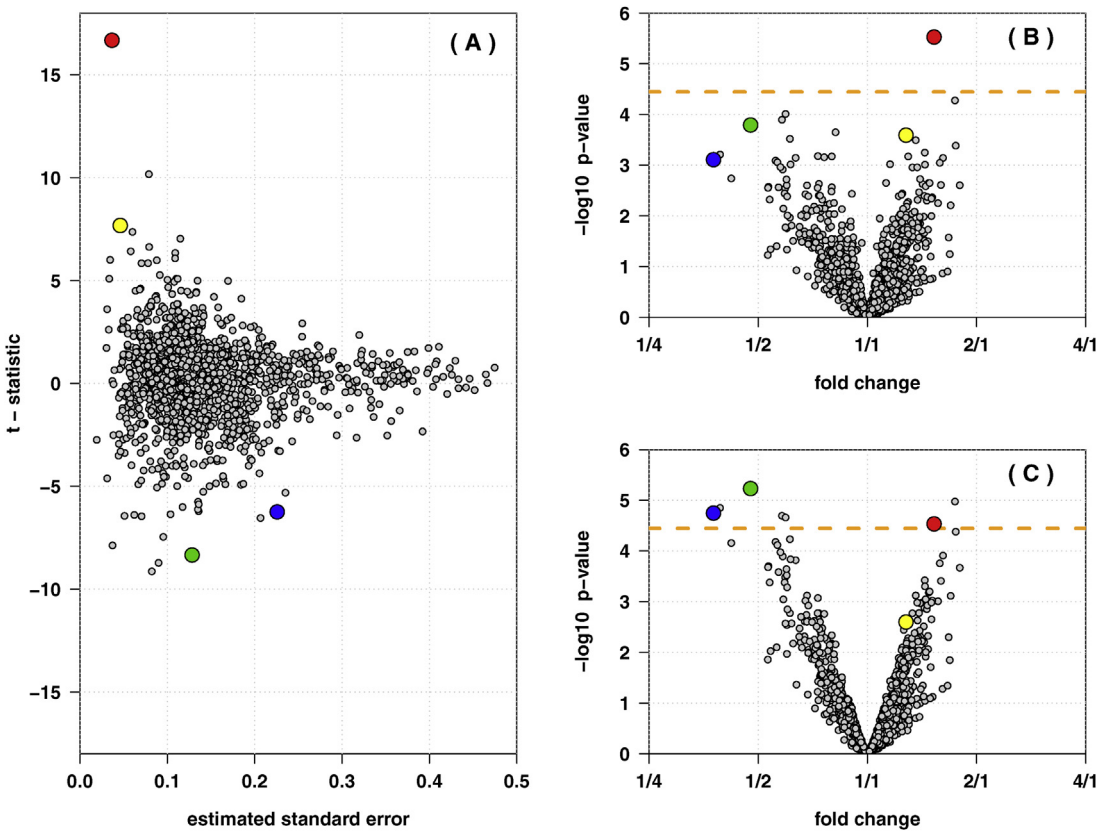

Fig. 1 – Inference from an 8-plex iTRAQ experiment with four cases and four controls, with 1394 proteins identified and quantified by 2 or more peptides. (A) The estimated standard error (*x*-axis) versus the *t*-statistics (*y*-axis) for each protein. The largest *t*-statistics (and thus, most significant *p*-values) tend to be from those proteins that show the smallest sample variability. (B) The volcano plot showing the estimated fold changes (*x*-axis) versus the $-\log_{10} p$-values (*y*-axis) for each protein. (C) The volcano plot from the inference based on the moderated *t*-statistics. Four proteins are highlighted in each panel, illustrating how proteins with small sample variability can show very low *p*-values despite small fold changes (red and yellow), while proteins with larger fold changes do not necessarily show significant differential expression (green and blue). The Bonferroni corrected significance level is indicated by the orange line. Pooling information from the entire distribution of all proteins improves power to detect differentially expressed proteins, and reduces statistical significance of proteins with small sample variability. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Bayes procedure LIMMA [13]. This approach shrinks the sample variances used in the estimation of the standard error of the observed fold changes towards a common mean. Thus, proteins with low sample variability have less significant moderated than ordinary *t*-statistics, and proteins with large sample variability have more significant moderated than ordinary *t*-statistics. Consequently, statistical significance increases for proteins with large fold change and relatively large sample variances (affecting the false negative rate) and statistical significance decreases for proteins with small fold change and relatively small sample variances (affecting the false positive rate). The observation that proteins with larger fold changes tend to become more significant (Supplementary Fig. 5) can be even more pronounced when proteins with larger fold changes also tend to have higher variability even after the logarithmic transformation of the relative abundances, i.e. when there is a mean-variance relationship (Supplementary Fig. 6). In the manuscript we have presented methods and results using isobaric mass labelled quantitative proteomic experiments, but note that these empirical Bayes methods are directly applicable in other settings as well. In fact, they can be used for any technology that yields measures of peptide or protein abundance, including label-free experiments.

We also showed how to jointly analyze data from multiple experiments, a necessity for example in iTRAQ experiments which limit the number of labels and thus the number samples that can be run simultaneously. Accounting for possible between experiment variability is important to maximize the power to detect differentially expressed proteins, and LIMMA provides a convenient and powerful framework to do so simply by specifying the design matrix (which also allows for experiments with multiple groups, beyond case–control studies [13]). An alternative approach to analyze data from multiple experiments, commonly used for example in meta analyses of genomic data from different studies, is based on inverse variance weighting (IVW): the observed effect sizes (fold changes) from different experiments are weighted

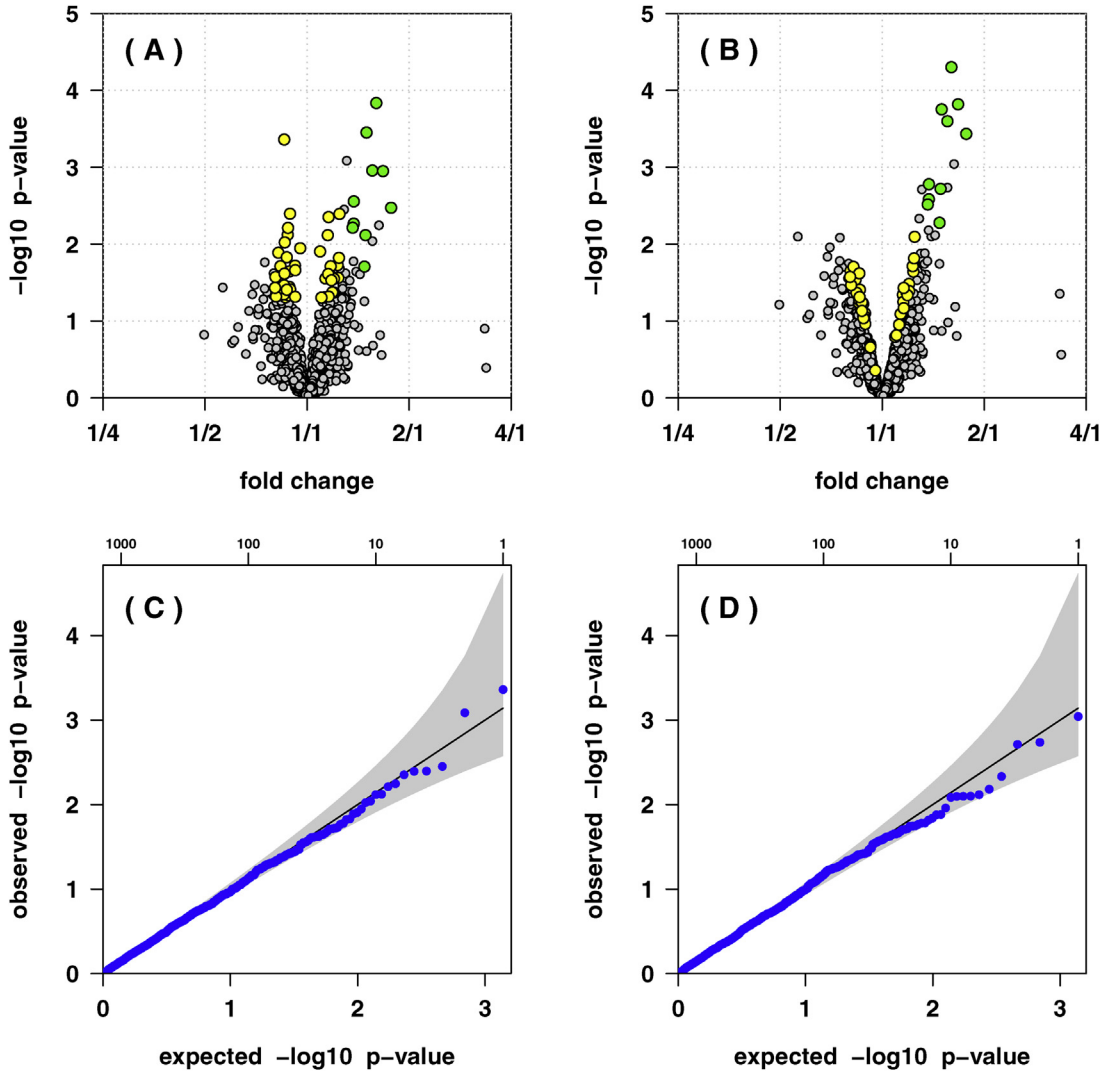

Fig. 2 – Inference for a simulated 8-plex iTRAQ experiment with four cases and four controls, and 1394 proteins. (A) The volcano plot showing the estimated fold changes (*x*-axis) versus the $-\log_{10}$ *p*-values (*y*-axis) for each protein. (B) The volcano plot from the inference based on the moderated *t*-statistics. The ten proteins simulated with a 50% fold change are highlighted in green. Pooling information from the entire distribution of all proteins improves power to detect these differentially expressed proteins. In addition, non-differentially expressed proteins with nominally significant *p*-values (less than 0.05) and small estimated fold changes (less than 25%) are highlighted in yellow, which highlights how pooling information from the entire distribution of all proteins reduces the false positive identification rate. The quantile–quantile (QQ) plots for the *p*-values of the non-differentially expressed proteins based on the ordinary 2-sample *t*-test (C) and the moderated *t*-test (D) show that both approaches properly control the type I error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by their respective inverse variances, and combined into a single test statistic. When sample sizes are such that the individual test statistics approximately follow a normal distribution, the IVW based test statistic also follows a normal distribution, and a *p*-value can readily be calculated. In proteomic experiments with data from few samples such as 8-plex iTRAQ, the test statistics follow *t*-distributions with only very few degrees of freedom, and thus are far from normality. In these settings the *p*-values have to be based on permutation tests (such as the one proposed in Storey et al. [37]) as the null distribution is unknown. In addition to the computational expense incurred, the procedure is also not very powerful in this setting as the weights themselves are subject to large variability due to the small sample size. We found that by first using empirical Bayes methods to moderate the variances in each experiment, and thus stabilizing the weights in the IVW procedure, the power to detect differentially expressed proteins can be greatly improved, although not up to the same level as the LIMMA model for the simultaneous analysis of multiple experiments (data not shown). This was even true for simulations where the prior scaling factor s_0^2 was doubled in one experiment compared to the other two. Additional complications using IVW can occur in the presence of missing data, and when sample sizes are extremely small,

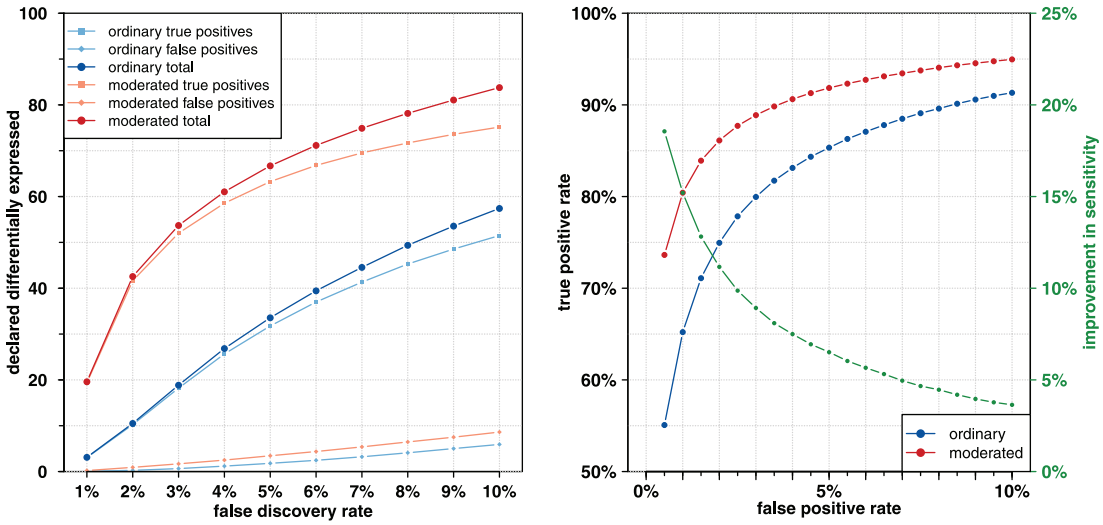

Fig. 3 – Results from the simulation study. Left: for all levels of false discovery rate control targeted (1–10%, *x*-axis), the empirical Bayes approach using moderated test statistic produces larger lists of proteins declared differentially expressed (*y*-axis), thus detecting more truly differentially expressed proteins while maintaining proper error control. Right: the same results presented in a ROC curve. For all false positive rates (0.5–10%, *x*-axis), the true positive rate (*y*-axis, black) is substantially higher when using moderated test statistics, yielding a large difference in sensitivity (*y*-axis, green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

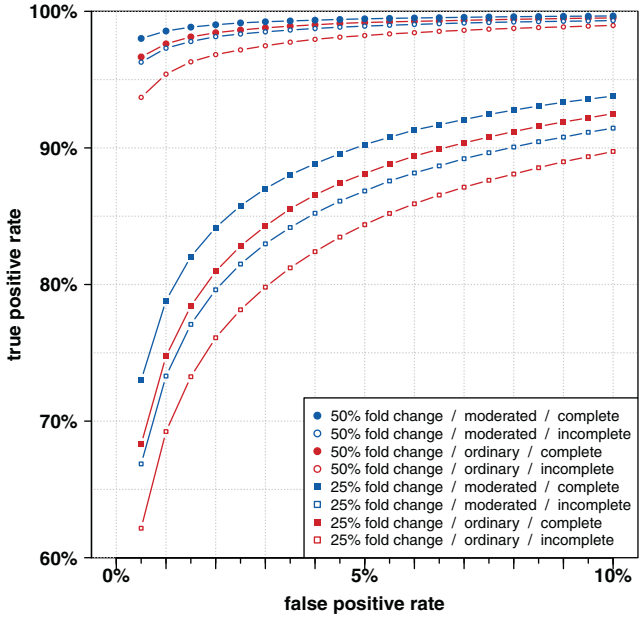

Fig. 4 – Results from the simultaneous analysis of three simulated iTRAQ experiments using the linear model in Eq. (8). For all false positive rates (0.5–10%, *x*-axis) considered, the true positive rate (*y*-axis) is substantially higher when using the moderated test statistics (blue) instead of the ordinary test statistics (red). Compared were fold changes of 50% (circles) and 25% (squares), with complete data (solid symbol) or 10% missing in each experiment (open symbol). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

as the actual number of possible permutations will be very low as well. Thus, LIMMA provides a more convenient and powerful framework to analyze such data than IVW.

The existence of missing data is largely owed to the nature of the methods used in proteomic studies, as peptide and protein identification between experiments can vary, sometimes substantially [38–41]. Although missing values for peptides and proteins due to biological differences are of scientific interests, these missing values can also be a nuisance arising from technical issues. Low or missing peptide intensities observed by the mass spectrometer can occur due to low peptide abundance, low peptide-dependent ionization efficiency, or the peptide signal distributed across multiple peptide charge states or different modified forms of the peptide. Peptide interactions with columns or other peptides in complex peptide mixtures or small differences in instrument ion sampling contribute to inconsistent peptide detection and missing values when comparing repeat MS analyses of the same sample. In multiplex analyses, such as isobaric mass tagging, missing values can also occur from signal dilution due to a low abundance peptide being present in one sample but a much lower abundance or absence in the other samples.

In this manuscript we also explained the effects of such missing data on the statistical inference with LIMMA, which performs a complete case analysis contrasting the cases to the controls within experiment. The amount of variance shrinkage depends on the actual observed experimental degrees of freedom – the more data (samples) are observed for a protein, the less the shrinkage towards the common mean variance. Missing data generally results in a loss of power to detect differentially expressed proteins, and depending on the nature of the missingness, can also introduce bias and affect the type I error [42,43]. In proteomic studies the missingness rate is commonly related to abundance [38,39]. However, in data

from multiple isobaric mass labelled proteomic experiments such as 8-plex iTRAQ we observe that the reporter ion intensity is usually observed in all channels within an experiment, or in none. That is, when a peptide is identified and quantified, the data are usually complete for all samples within an experiment [11]. Thus, the rate of missingness is much stronger related to the sampling method used by the instrument (i.e., experiment ID) than absolute abundance, let alone relative abundance (fold change). For other types of experiments such as label-free quantification it might be necessary however to address the missing data explicitly for example using imputation methods [44–46] before employing empirical Bayes methods, since the missingness pattern might be strongly related to protein abundance. This would violate the notion of "missingness completely at random", an implicit assumption in complete case analyses [42,43].

Conflict of interest

The authors declare no conflict of interest.

Transparency document

The Transparency document associated with this article can be found in the online version.

Acknowledgments

Support was provided by the Deutsche Forschungsgemeinschaft (KA 3884/1-1) and the National Institutes of Health (AI058613). The contributions of Robert E. Jensen are gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.euprot.2015.02.002.

REFERENCES

[1] Urfer W, Grzegorczyk M, Jung K. Statistics for proteomics: a review of tools for analyzing experimental data. Proteomics 2006;6(Suppl. 2):48–55.
[2] Keshamouni VG, Michailidis G, Grasso CS, Anthwal S, Strahler JR, Walker A, et al. Differential protein expression profiling by iTRAQ-2DLC–MS/MS of lung cancer cells undergoing epithelial–mesenchymal transition reveals a migratory/invasive phenotype. J Proteome Res 2006;5(5):1143–54.
[3] Gan CS, Chong PK, Pham TK, Wright PC. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). J Proteome Res 2007;6(2):821–7.
[4] Prakash A, Piening B, Whiteaker J, Zhang H, Shaffer SA, Martin D, et al. Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics. Mol Cell Proteomics 2007;6(10):1741–8.
[5] Vitek O. Getting started in computational mass spectrometry-based proteomics. PLoS Comput Biol 2009;5(5):e1000366.
[6] Kaell L, Vitek O. Computational mass spectrometry-based proteomics. PLoS Comput Biol 2011;7(12):e1002277.
[7] Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, et al. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. J Proteome Res 2008;7(1):225–33.
[8] Hill EG, Schwacke JH, Comte-Walters S, Slate EH, Oberg AL, Eckel-Passow JE, et al. A statistical model for iTRAQ data analysis. J Proteome Res 2008;7(8):3091–101.
[9] Kathleen Kerr M, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol 2000;7(6):819–37.
[10] Kathleen Kerr M, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001;2(2):183–201.
[11] Herbrich SM, Cole RN, West KP, Schulze K, Yager JD, Groopman JD, et al. Statistical inference from multiple iTRAQ experiments without using common reference standards. J Proteome Res 2013;12(2):594–604.
[12] Box GEP, Draper NR. Empirical model-building and response surfaces. New York, NY: John Wiley & Sons; 1987. p. 424.
[13] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3(1).
[14] Brusniak M-Y, Bodenmiller B, Campbell D, Cooke K, Eddes J, Garbutt A, et al. Corra: computational framework and tools for LC–MS discovery and targeted mass spectrometry-based proteomics. BMC Bioinform 2008;9:542.
[15] Ting L, Cowley MJ, Hoon SL, Guilhaus M, Raftery MJ, Cavicchioli R. Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. Mol Cell Proteomics 2009;8(10):2227–42.
[16] Margolin AA, Ong S-E, Schenone M, Gould R, Schreiber SL, Carr SA, et al. Empirical Bayes analysis of quantitative proteomics experiments. PLoS ONE 2009;4(10):e7454.
[17] Jankova L, Chan C, Fung CLS, Song X, Kwun SY, Cowley MJ, et al. Proteomic comparison of colorectal tumours and non-neoplastic mucosa from paired patient samples using iTRAQ mass spectrometry. Mol Biosyst 2011;7(11): 2997–3005.
[18] Schwaemmle V, Leon IR, Jensen ON. Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates. J Proteome Res 2013;12(9):3874–83.
[19] Zhao S, Li R, Cai X, Chen W, Li Q, Xing T, et al. The application of SILAC mouse in human body fluid proteomics analysis reveals protein patterns associated with IgA nephropathy. Evid Based Complement Altern Med 2013;2013:275390.
[20] Schwacke JH, Hill EG, Krug EL, Comte-Walters S, Schey KL. iQuantitator: a tool for protein expression inference using iTRAQ. BMC Bioinform 2009;10:342.
[21] Breitwieser FP, Mueller A, Dayon L, Koecher T, Hainard A, Pichler P, et al. General statistical modeling of data from protein relative expression isobaric tags. J Proteome Res 2011;10(6):2758–66.
[22] Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 2012;30(10):918–20.
[23] Wang P, Yang P, Yang JYH. OCAP: an open comprehensive analysis pipeline for iTRAQ. Bioinformatics 2012;28(10):1404–5.
[24] Gatto L, Lilley KS. MSnbase – an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics 2012;28(2):288–9.

[25] Gatto L, Christoforou A. Using R and Bioconductor for proteomics data analysis. Biochim Biophys Acta 2014;1844(1 Pt A):42–51.
[26] Choi M, Chang C, Clough T, Broudy D, Killeen T, MacLean B, et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 2014;30(17):2524–6.
[27] Li Z, Lindsay ME, Motyka SA, Englund PT, Wang CC. Identification of a bacterial-like HslVU protease in the mitochondria of *Trypanosoma brucei* and its role in mitochondrial DNA replication. PLoS Pathog 2008;4(4):e1000048.
[28] Liu B, Wang J, Yaffe N, Lindsay ME, Zhao Z, Zick A, et al. Trypanosomes have six mitochondrial DNA helicases with one controlling kinetoplast maxicircle replication. Mol Cell 2009;35(4):490–501.
[29] Ringpis G-E, Lathrop RH, Aphasizhev R. iCODA: RNAi-based inducible knock-in system in *Trypanosoma brucei*. Methods Mol Biol 2011;718:23–37.
[30] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3(12):1154–69.
[31] Pierce A, Unwin RD, Evans CA, Griffiths S, Carney L, Zhang L, et al. Eight-channel iTRAQ enables comparison of the activity of six leukemogenic tyrosine kinases. Mol Cell Proteomics 2008;7(5):853–63.
[32] Wang Y, Yang F, Gritsenko MA, Wang Y, Clauss T, Liu T, et al. Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. Proteomics 2011;11(10):2019–26.
[33] Rice J. Mathematical statistics and data analysis. 2nd ed. Duxbury Press; 1995.
[34] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57(1):289–300.
[35] Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B 2002;64(3):479–98.
[36] Storey JD. The positive false discovery rate: a Bayesian interpretation and the *q*-value. Ann Stat 2003;31(6):2013–35.
[37] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003;100(16):9440–5.
[38] Liu H, Sadygov RG, Yates JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 2004;76(14):4193–201.
[39] Wang P, Tang H, Zhang H, Whiteaker J, Paulovich AG, Mcintosh M. Normalization regarding non-random missing values in high-throughput mass spectrometry data. Pac Symp Biocomput 2006;31:5–326.
[40] Chong PK, Gan CS, Pham TK, Wright PC. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: implication of multiple injections. J Proteome Res 2006;5(5):1232–40.
[41] Jung K, Dihazi H, Bibi A, Dihazi GH, Beissbarth T. Adaption of the global test idea to proteomics data with missing values. Bioinformatics 2014;30(10):1424–30.
[42] Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91(434):473–89.
[43] Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8(1):3–15.
[44] Pedreschi R, Hertog MLATM, Carpentier SC, Lammertyn J, Robben J, Noben J-P, et al. Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. Proteomics 2008;8(7):1371–83.
[45] Albrecht D, Kniemeyer O, Brakhage AA, Guthke R. Missing values in gel-based proteomics. Proteomics 2010;10(6):1202–11.
[46] Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC–MS analysis. BMC Bioinform 2012;16(13 Suppl.):S5.